DOCUMENT RESUME

ED 041 932 TM 000 016

AUTHOR Shoemaker, David M.

TITLE Allocation of Items and Examinees in Estimating a Norm Distribution by Item Sampling.

PUB DATE Mar 70

NOTE 10p.; Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minn., March 1970. Accepted for publication in the Journal of Educational Measurement

EDRS PRICE EDRS Price MF-$0.25 HC-$0.60

DESCRIPTORS Item Sampling, *Multiple Choice Tests, *National Norms, Norms, *Sampling, *Statistical Analysis

ABSTRACT

A norm distribution consisting of test scores received by 810 college students on a 150 item dichotomously-scored four-alternative multiple-choice test was empirically estimated through several item-examinee sampling procedures. The post mortem item-sampling investigation was specifically designed to manipulate systematically the variables of numbe: of subtests, number of items per subtest, and number of examinees responding to each subtest. Defining one observation as the score received by one examinee on one item, the results suggest that as the number of observations increases beyond 1.23 per cent of the data base all procedures produce stochastically equivalent results. The results of this investigation indicate that, in estimating a norm distribution by item-sampling, the variable of importance is not the item-sampling procedure per se but is instead the number of observations obtained by the procedure. It should be noted, however, that in this investigation the test score norm distribution was approximately symmetrical and the possibility should not be overlooked that item-sampling as a procedure may be robust for symmetrical norm distributions. (Author)

ALLOCATION OF ITEMS AND EXAMINEES IN ESTIMATING A NORM DISTRIBUTION BY ITEM-SAMPLING

David M. Shoemaker

Southwest Regional Laboratory
for
Educational Research and Development

The negative hypergeometric distribution has been found to provide a reasonably good fit for a variety of test score distributions where the test score is the number of correct answers (Keats & Lord, 1962). The negative hypergeometric distribution is, within this context, a function of three parameters: the number of test items and estimates of the mean and variance of the normative distribution. Operating within the framework of the item-sampling model, Lord (1960) has provided the appropriate equations for computing unbiased estimates of the first two moments of a frequency distribution and has, furthermore, demonstrated (Lord, 1962) that a norm distribution may be satisfactorily approximated by a negative hypergeometric distribution fitted to parameters estimated through item-sampling. The procedure is as follows:

1. The test items to be normed are divided into $\underline{t}$ subtests and each subtest is administered to a different set of examinees.
2. The results obtained from each subtest (item-examinee sample) provide an estimate of the mean μ and variance σ^2 of the norm distribution when formulas 9 and 10 in Lord (1962) are applied. A single estimate of μ is obtained by averaging the $\underline{t}$ estimates of μ obtained from each item-examinee sample; a single estimate of σ^2, by averaging the $\underline{t}$ estimates of the population variance.

3. Substituting each possible test score x into the negative hypergeometric function specified in equation 23.6.10 in Lord and Novick (1968) produces an estimate of the proportion of examinees in the norm population receiving that test score.[1]

Implementing the procedure outlined above produces many interesting questions: How many different subtests _t_ of items and examinees are required to estimate satisfactorily the norm distribution? Is it more appropriate to administer a fewer number of subtests containing a larger number of items or a larger number of subtests containing fewer items? To how many examinees should each subtest be administered? Must all items in the test be distributed among the subtests? The project described herein was an attempt to provide tentative answers to questions such as these.

Several investigations (e.g., Plumlee, 1964; Cahen _et al_., 1969; Owens & Stufflebeam, 1969) have estimated parameters by item-sampling but only Cook and Stufflebeam (1967) have investigated the relative merits of different item-sampling procedures in estimating a norm distribution with the negative hypergeometric distribution. It should be mentioned that the expressed purpose of their study was that of contrasting two approaches — item sampling, given the condition of sampling without replacement, and examinee sampling — in estimating a norm distribution. Cook and Stufflebeam concluded that item sampling is equally effective to examinee sampling.

In the Cook and Stufflebeam (1967) design, the number of subtests is confounded with number of items per subtest and with number of examinees receiving each subtest. Using the Cook and Stufflebeam article as a point of departure, the present investigation was specifically designed to manipulate systematically the variables of number of subtests, number of items per subtest, and number of examinees responding to each subtest to determine the relative merits of several item-sampling procedures which might be used in estimating a norm distribution.

METHOD

The research design was one of *a posteriori* item-sampling: given a norm distribution, various item-examinee samples are selected at random from this data base and used to estimate the distribution from which they have been sampled. In this investigation the norm distribution consisted of test scores received by 810 college students on a 150 item dichotomously-scored 4—alternative multiple-choice test administered as a final examination in the Spring of 1969. On this examination the mean score μ was 87.390 with variance σ^2 of 324.193 and Kuder-Richardson Formula 21 reliability equal to .893.

The twenty item-sampling procedures used to estimate the norm distribution are listed in Table 1. As all procedures, with one exception, are similar only procedure 1 will be described in detail.

Please insert Table 1 about here.

In procedure 1, the 150 test items were divided by randomly sampling without replacement into 10 subtests each containing 15 items. From the pool of 810 examinees, 10 groups of 10 examinees were selected at random and without replacement. Each subgroup was administered one subtest, that is, only those items in that subtest were scored for those examinees.[2] Procedure 1 produced 10 estimates of μ and σ^2. The pooled estimate of μ was found to be 87.111; the standard deviation of the 10 estimates of μ was 14.867. The pooled estimate of σ^2 was 318.185 and the standard deviation of the 10 estimates was 263.033. Using these estimates of the parameters, the Kuder-Richardson Formula 21 reliability coefficient for the full-length test was computed to be .891. The absolute value of the maximum difference D_{max} between the cumulative relative negative hypergeometric distribution fitted to the estimates of μ and σ^2 obtained from procedure 1 and the cumulative relative negative hypergeometric distribution fitted to μ and σ^2 was .038. D_{max} between each pair of distributions, the test statistic for the Kolmogorov-Smirnov one-sample test for goodness-of-fit (Siegel, 1954), was selected from 150 differences.

Procedures 1 through 4 are similar to the item-sampling procedures used by Cook and Stufflebeam (1967) with the exception that the number of examinees receiving each subtest has been held constant. Procedures 5 through 8 are a replication of 1 through 4 with an increase in the number of examinees receiving each subtest. In procedures 9 through 12 the number of items per subtest and the number of examinees receiving each subtest have been held constant; in 13 through 16, the number of subtests and the number of examinees receiving each subtest have been held constant. In procedures 17 through 20 only the number of examinees receiving each subtest has been held constant.

Each set of four procedures was a systematic exploration of the Cook and Stufflebeam (1967) design. Certain procedures, i.e., 1 and 9, 2 and 13, 10 and 14, 2 and 18, 12 and 20, are identical and were computed once; in each instance the results were recorded twice in Table 2.

RESULTS AND DISCUSSION

All results are recorded in Table 1. On the basis of the Kolmogorov-Smirnov one-sample test[3], three procedures produced negative hypergeometric distributions which were judged not to be stochastically equivalent[4] to the fitted norms distribution. In Procedures 1 through 4, with the number of examinees per subtest being held constant, all negative hypergeometric distributions were equivalent to the fitted norms distribution. While it is of theoretical interest to note that the smallest value of D_{max} occurred with that item sampling procedure involving a large number of subtests with few items per subtest--with the converse being also true, the effect was nullified (procedures 5 through 8) by an increase in the number of examinees receiving each subtest. Procedures 9 through 16 were designed to partial out the effect noted in procedures 1 through 4. Holding the number of items per subtest and the number of examinees per subtest constant, an increase in the number of subtests produced a negative hypergeometric distribution more stochastically equivalent to the fitted norms distribution. Similar results were obtained (procedures 13 through 16) with an increase in the number of items per subtest, holding constant the number of subtests and number of examinees per subtest. The results from procedures 17 through 20 suggest that beyond a certain point little is to be gained by simultaneously increasing the number of subtests and the number of items per subtest.

The inconsistencies found in Table 1 (e. g., procedures 17 and 20 producing negative hypergeometric distributions equivalent to the fitted norms distribution) are made less alarming if D_{max} per procedure is analyzed as a function of the number of observations (one observation is equal to the score received by one examinee on one item). For small numbers of observations the values of D_{max} are variable and inconsistent; however, as the number of observations increases beyond a certain point, all procedures produce equivalent results. That certain point in this investigation was approximately 1.23% of the norm data base. It is not surprising, therefore, that Lord (1962) and Plumlee (1964) obtained a good approximation with an item-sampling procedure involving 10% of the total observations and that similar results were obtained by Cook and Stufflebeam (1967) with procedures involving percentages of total observations ranging from 9.18 to 49.45.

REFERENCES

Cahen, L. S., Romberg, T. A. and Zwirner, W. The estimation of mean achievement scores for schools by the item-sampling technique. Educational and Psychological Measurement, 1969 (in press).

Cook, D. L. and Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. Educational and Psychological Measurement, 1967, 27, 601-610.

Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Lord, F. M. Use of true-score theory to predict moments of univariate and bivariate observed-score distributions. Psychometrika, 1960, 25, 325-342.

Lord, F. M. Estimating norms by item-sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Owens, T. R. and Stufflebeam, D. L. An experimental comparison of item-sampling and examinee sampling for estimating test norms. Journal of Educational Measurement, 1969, 6, 75-83.

Plumlee, Lynnette B. Estimating means and standard deviations from partial data--an empirical check on Lord's item-sampling technique. Educational and Psychological Measurement, 1964, 24, 623-630.

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

TABLE 1

Item sampling procedures with results

	Procedure					Results					
	No. of sub-tests	No. of items per subtest	No. of examinees per subtest	Total observa-tions	Total N	$\hat{\mu}$	$\hat{\sigma}^2$	SD($\hat{\mu}$)	SD($\hat{\sigma}^2$)	KR21	D_{max}
(1)	10	15	10	1500	100	87.111	318.185	14.867	263.033	.891	.038
(2)	5	30	10	1500	50	88.700	258.824	12.087	103.999	.866	.051
(3)	3	50	10	1500	30	90.600	209.519	4.655	104.849	.834	.085
(4)	2	75	10	1500	20	86.700	223.516	12.100	140.523	.842	.100
(5)	10	15	20	3000	200	88.850	307.578	11.030	181.854	.888	.038
(6)	5	30	20	3000	100	88.050	327.146	4.299	172.028	.895	.000
(7)	3	50	20	3000	60	87.550	327.210	4.848	188.019	.895	.000
(8)	2	75	20	3000	40	88.150	310.004	.499	8.703	.889	.038
(9)	10	15	10	1500	100	87.111	318.185	14.867	263.033	.891	.038
(10)	5	15	10	750	50	91.800	508.148	9.848	496.893	.936	.089
(11)	3	15	10	450	30	101.333	450.334	4.498	180.526	.933	.350*
(12)	2	15	10	300	20	94.500	348.410	8.500	14.792	.906	.120
(13)	5	30	10	1500	50	88.700	258.824	12.087	103.999	.866	.051
(14)	5	15	10	750	50	91.800	508.148	9.848	496.893	.936	.089
(15)	5	10	10	500	50	78.600	328.090	18.288	94.095	.892	.204*
(16)	5	5	10	250	50	75.000	375.765	6.481	44.909	.906	.274*
(17)	10	40	10	4000	100	90.041	415.972	9.383	161.014	.920	.100
(18)	5	30	10	1500	50	88.700	258.824	12.087	103.999	.866	.051
(19)	3	20	10	600	30	87.999	396.823	.945	241.973	.914	.060
(20)	2	15	10	300	20	94.500	348.410	8.500	14.792	.906	.120
Norm	1	150	810	121500	810	87.390	324.193			.893	

*Negative hypergeometric distribution not stochastically equivalent to fitted norms distribution.

FOOTNOTES

[1] In the computer program used for calculating values of this proportion, 1/2 was added to _a_ and _b_ as defined in Lord and Novick (1968). Each term was truncated before substitution into equation 23.6.10. A copy of the Fortran program with documentation may be obtained upon request from the author.

[2] The exception to this general pattern was found in procedure 17. Each subtest was formed by randomly sampling without replacement 40 items from the 150 item pool. It was, therefore, possible for a particular item to appear in more than one subtest. Only 2 of the 150 items were not selected for inclusion in any subtest.

[3] A referee has pointed out, and correctly so, that the Kolmogorov-Smirnov tests should be viewed as providing rough indications rather than strict significance tests. Since the population was finite and since the sampling was done without replacement, there is necessarily a closer agreement between sample and population than there would be in random sampling from an infinite population.

[4] Two distributions are said to be stochastically equivalent if the two distributions are distinct and if $f(x) = g(x)$ for all x.

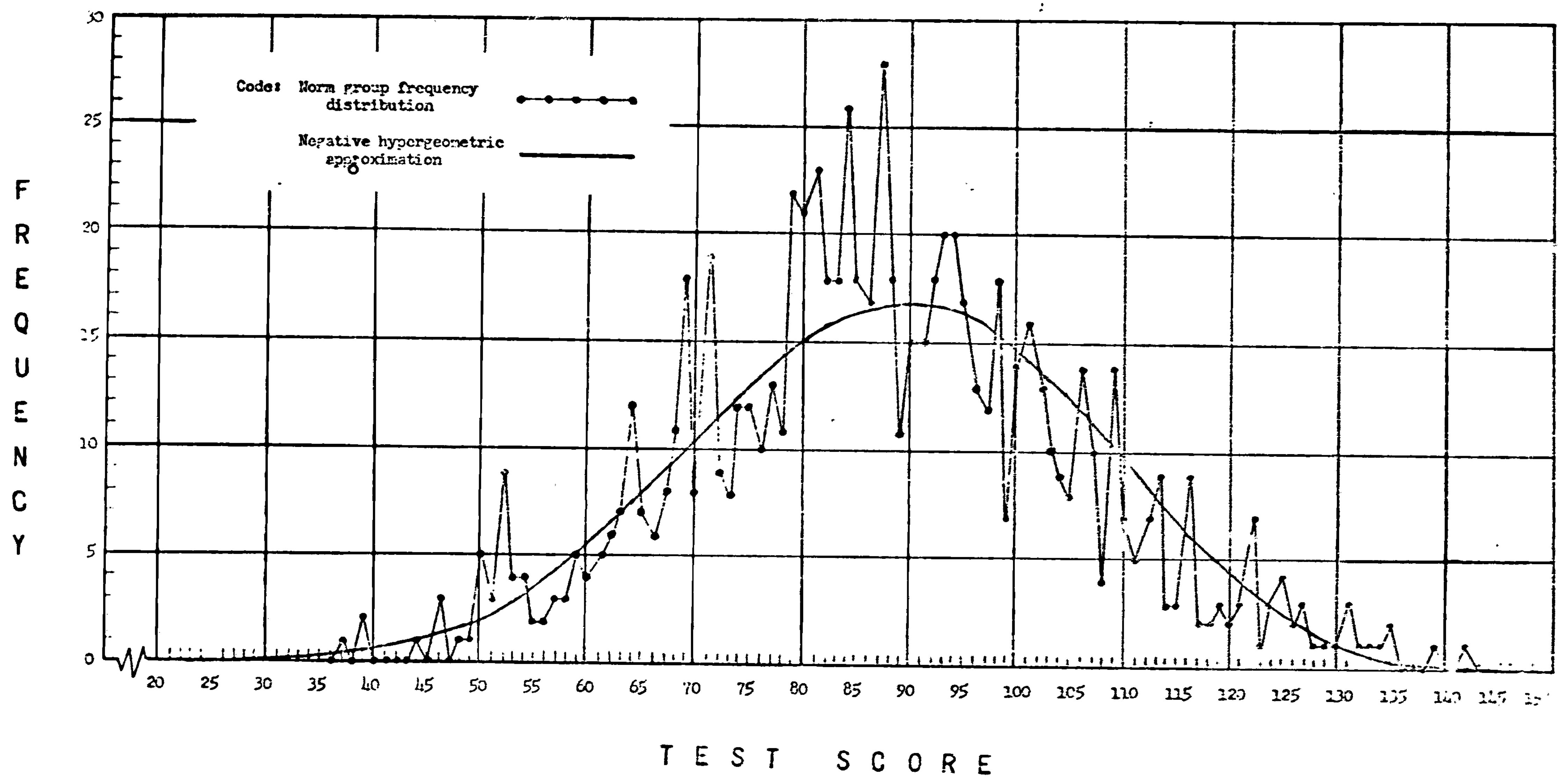

Figure 1. Norms group frequency distribution and negative hypergeometric distribution fitted to the parameters. (Frequency polygons were used for graphic clarity. Both distributions represent discrete, not continuous, variables.)